ISSN 2277-2685
IJESR/Sep. 2023/ Vol-13/Issue-3/1-10
Mrs.S.Swapna *et. al.*, / International Journal of Engineering & Science Research

WELFake: Word Embedding Over Linguistic Features for Fake News Detection

Mrs.S.Swapna [1], Ileni Ashritha [2], N Abhilash [2], Ajmeera Sravan [2], Kola Omprasad [2]
[1]*Asst. Professor,Computer Science and Engineering, CMR Engineering College, medchal, T.S, India*
[2]*B.Tech, Computer Science and Engineering, CMR Engineering College, medchal, T.S, India*

Abstract: Real-time news is widely disseminated via social media on a global scale. One of the factors contributing to its success is the simple and speedy spread of information. Social networking platforms have a huge user base that includes people of all ages, genders, and social backgrounds. Despite these positive characteristics, a serious drawback is false news, since most individuals read and spread information without giving any thought to its veracity. Researching techniques for news authenticity is so essential. This article suggests a two-phase benchmark model for detecting fake news using machine learning classification called WELFake, which is built on word embedding (WE) over linguistic variables. The data set is preprocessed in the first stage, and linguistic characteristics are used to verify the accuracy of the news material. The second stage conducts voting classification while combining the linguistic feature sets with WE. This paper meticulously creates a unique WELFake data set with over 72 000 articles to verify its methodology. This data set combines many data sets to provide an objective classification result. According to experimental findings, the WELFake model classifies news as genuine or false with a 96.73% accuracy rate, which is better than convolutional neural network (CNN) models and bidirectional encoder representations from transformer (BERT) models by 1.31% and 4.25%, respectively. The predictive-based comparable works created using the Word2vec WE approach are outperformed by our frequency-based and targeted analysis of writing patterns model by up to 1.73%.

1. INTRODUCTION

Nowadaysno of their age, community, or sex, individuals are becoming more interested on online social networks [1]. Using social networks for communication makes it easy, quick, and appealing to exchange and transmit information. The industry leaders right now are social networking sites like Facebook and Twitter, which serve over 1.3 billion users on a monthly average fluctuation of 300 million people [2]. Terabytes of data are produced every second as a result of their partnerships [3], [4]. Online social networks are appealing due to how easy and practical it is to obtain and share information with others. However, the broad dissemination of erroneous information, such as fake news, which is destructive to society and individuals, is made possible by the quick dispersal of data at a high pace with little effort.

Fake news is low-quality material that contains willfully misleading information and is spread by people or automated programmes that actively distort messages for gossip or political purposes. According to Schudson and Zelizer [5], the phrase "fake news" and the mass media itself both date back to earlier ages.

However, during the 2016 U.S. presidential elections, when the spread of false news on social media captured the attention of more internet users than conventional newsreaders, this phrase gained more traction. A link to a very biassed or misleading news website was included in almost 7.5 million tweets in the last five months before the elections. The fact that misleading and unsupported news from dubious sources draws larger viewers than accurate information is both intriguing and alarming [6]. Relevant research on the subject has shown that false news has a greater effect than actual news in terms of speed of distribution and depth of penetration [7]. People often accept and share news without verifying its accuracy or the reliability of its sources. By doing this, people join a group that mistakenly or purposely spreads false information. It's possible that the spread of false news is motivated by the desire to amuse oneself or to influence public opinion for financial or political gain. The bad effects of this phenomena are clearly evident and include anything from poor judgement to instances of bullying and violence.

A Description Of The Project: There is a critical need for methods to check the validity of the material since online social networks are important sources of information that might deceive people or groups . In an effort to automate the false news identification process, several researchers continue to work to construct machine learning (ML) models using a variety of feature sets employing visual or text-based linguistic techniques. The next four queries, however, are still unaddressed.

1) What language characteristics are most important for separating the true news from the fake?

2) When compared to other ML techniques like convolutional neural networks (CNNs) or bidirectional encoder representations from transformers (BERTs) , which word embedding (WE) methodology using linguistic data predicts false news the best?

3) Which classification technique is best for identifying false news in the data sets at hand?

4) Does the ensemble voting classifier enhance the outcomes of the false news detection?

We provide a novel, three-stage approach dubbed WELFake that is only targeted at text data to address these inquiries.

1) Linguistic feature sets (LFS) are used to forecast false news; 2) WE over LFS is used to identify fake news more effectively than WELFake data.

3) A comparison of the linguistic feature-based outcomes using cutting-edge CNN and BERT techniques.

For the classification of authentic and false news, the WELFake model does not need any extra metadata information pertaining to the user or media . Instead, it uses a combined LFS and WE strategy to modernise state-of-the-art methods for identifying false news on social networking networks. Three contributions made by our WELFake model are highlighted.

2. LITERATURE SURVEY

2.1 Existing System

There is a critical need for methods to check the validity of the material since online social networks are important sources of information that might deceive people or groups . In an effort to automate the false news identification process, several researchers continue to work to construct machine learning (ML) models using a variety of feature sets employing visual or text-based linguistic techniques.[8]

2.2Proposed System

The system suggests a novel, three-stage approach dubbed WELFake that is only targeted at text data. 1) Linguistic feature sets (LFS) are used to forecast false news; 2) WE over LFS is used to identify fake news more effectively than WELFake data.

3) A comparison of the linguistic feature-based outcomes using cutting-edge CNN and BERT techniques. For the categorization of authentic and false news, the WELFake model does not need any extra metadata information pertaining to the user or media . Instead, it uses a combined LFS and WE strategy to modernise state-of-the-art methods for identifying false news on social networking networks. Three contributions made by our WELFake model are highlighted.

2.2 Methods for ML Classification

In this part, we examine a few ML techniques utilised in the WELFake model to identify bogus news.

1) Naive Bayes: This supervised learning method is based on Bayes' theorem and provides quick predictions with improved accuracy in the areas of sentiment analysis, spam filtering, and text categorization.

2) Support Vector Machine: This supervised learning method solves classification and regression issues simultaneously. The programme predicts the right set for fresh data values and determines the optimum line for set separation.

3) Decision Tree: This supervised learning approach divides the data into categories for both continuous and categorical dependent variables. This classifier divides entire data sets into homogenous ones in order to solve an issue. The data set, the decision criteria, and the result are represented in this tree structure's internal nodes, branches, and leaf nodes. Information gain and Gini index are two attribute selection metrics for the optimal attribute node.

4) Random Forest: Based on ensemble learning, this supervised learning method ensembles many decision trees (DTs) into a random forest (RF) and computes the average outcomes. The accuracy of the model might be improved by the high number of trees in the RF.

5) K-Nearest Neighbour is a good algorithm for classification issues based on feature similarity. Based on the problem description and statistics, the method may utilise any integer value for K, and it uses the Euclidean, Manhattan, or Hamming metric to determine the separation of the data.

6) Boosting: This links each base learner in turn.It initially sends a small number of records to the first base learner (B L1) of any model for training, then it assesses every record on B L1 and sends the records that were improperly categorised to the second learner (B L2). All records are examined by B L2 before being passed on to B L3 for classification. The procedure keeps on until a certain number of basic learners have been reached.

7) Bagging: Also referred to as bootstrap aggregation, this ensemble strategy employs several base learners and gives each model a distinct portion of the original data set for training. Based on the aggregated votes from the various models, the testing procedure selects the outcome. In addition to diverse sample sets, different subsets may be used to train the models to minimise over-fitting.[9]

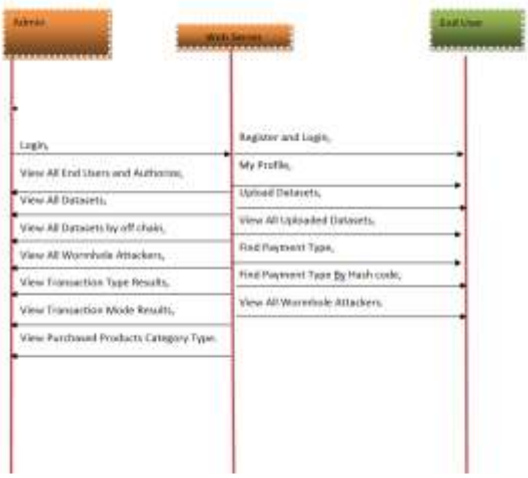

Fig 1: System Flow

3. SYSTEM ANALYSIS&DESIGN

3.1 Clickbaits

Clickbaits are a specific kind of bogus material that uses catchy headlines to draw readers in but then falls short of delivering on those promises. By merging syntax and semantics under textual cues and visuals with newsreader behaviour under nontextual signals, Chen et al. examined possible strategies for programed detection of clickbaits. They examined the significance of these indicators for spotting false news, but they were unable to put their research into practise.

A methodology for clickbait identification that evaluates the relevance of headlines for article bodies was put out by Bourgonje et al. [10].

Using the data set made available by the inaugural fake news challenge organisers, they were able to identify stances with a considerable accuracy of 89.59% by using a logistic regression classifier. Rashkin et al. identified the characteristics of false text and other internet sources for fact-checking by contrasting the language of true news with that of satire, hoaxes, and propaganda.

B. News Reputation

Alrubaian et al. used text-specific variables (sentiment analysis), user-specific features (gender, number of followers), and message-specific data (URL, hashtag, number of responses) to analyse the trustworthiness of 489 330 Twitter accounts. With tenfold cross-validation, they built the RF, Naive Bayes (NB), DT, and feature-rank NB algorithms. Castillo and colleagues identified fake news based on credibility using four feature sets: text-specific (e.g., number of characters, words, question marks, punctuation, and sentiment analysis); user-specific (e.g., verified accounts, number of followers, number of tweets, and account creation time); message-specific (e.g., with URL, retweeted status); and propagation (e.g., root node degree and propagation tree depth).[11]

Precision and recall for these characteristics on the Twitter data set were between 70% and 80%. According on stylistic (such as syntax, text style, and grammar), complexity (such as article title), and psycholinguistic aspects, Benjamin et al. rated the news as authentic or fraudulent. Larger volumes of data and more detailed characteristics provide better outcomes, according to experimental findings on three different data sets. On the Kaggle data set, Kaliyar et al.'s deep convolutional neural network for false news detection network (FNDNet) model obtained a 98.36% accuracy. They did not, however, display any outcomes for generalised text . On the PolitiFact and GossipCop data sets, Shu et al.'s discussion of the accuracy of real-news and fake-news detection via comments on specific articles yielded 90.4%, 80.80%, and 80.80% accuracy, respectively. A text generation approach called GROVER, put out by Zellers et al. , employs opponents rather than people to disseminate false information that is more reliable. They looked at a number of threats that were spread, demonstrating that GROVER may be a powerful discriminator that surpasses BERT in spotting fake news.

3.2 Language Features

Since the middle of the 2000s, linguistic characteristics have been widely used for false news identification. Burgoon et al. employed a DT algorithm with 15-fold cross-validation to reach an accuracy of 60.72% utilising 16 linguistic characteristics divided into four groups. Vicario et al. identified hoaxes and fake news on social media using linear regression, logistic regression, support vector machine (SVM), K-nearest neighbour (KNN), and NNs. These features included text (e.g., number of characters, words, sentences, question marks, and negations), user-specific, and message-specific (e.g., number of replies, likes). The validation on a new feature-rich Italian Facebook data set produced a linear regression classification algorithm accuracy of 91%. On two new data sets encompassing seven categories, Pérez-Rosas et al. utilised significant linguistic variables (such as n-grams, punctuation, psycho-linguistic, readability, and syntax) and reached an accuracy of 76%. Using 45 features from the following four categories—structural (e.g., length, number of tweets), content (e.g., polarity, subjectivity), temporal, and user (e.g., age, followers, authenticity)—Buntain and Golbeck tried to categorise actual or false Twitter threads. They achieved an accuracy of 65.29% using the BuzzFeed data set for testing and the CREDBANK and PHEME data sets for training. Using logistic regression, Newman et al. evaluated 29 characteristics with an accuracy ranging from 52% to 67%. Similar to this, Zhou et al. used 20 variables across nine categories to identify bogus news.

Ahmed et al.'s accuracy for detecting false news was 92% using an ambiguous private version of the continuously updated Kaggle data set. Politifact and BuzzFeed data sets, which are very tiny and difficult to generalise, were employed by Shu et al. in their research. By merging select articles from four data sets (Kaggle-EXT, McIntire, BuzzFeed, and Politifact) in a new UNBiased data set containing 3004 articles, Gravanis et al. compared their model with and obtained the greatest accuracy of 95% using SVM. Five kinds of linguistic features are compared in Table I.

1) The readability index measures the complexity (i.e., reading difficulty) of the text using the word length, syllable count, and sentence length.

2) Emotions, behaviours, personas, and attitude are described by psycho-linguistic characteristics.

3) Stylistic elements describe a sentence's style.

4) User information is described by user trustworthiness characteristics.

5) Quantity characteristics provide information on sentence quantity, such as the amount of words and phrases.

3.3 System Architecture:Prepare a data set

1) News Data Collection: This is the first step in creating a fair and objective data collection, as well as the foundation for producing high-quality training data and successful outcomes. Although there are several open data sets available for the research of false news (see Section III and Table II), literature has shown that these sets have major size, category, or bias constraints.

We produced a more thorough WELFake data set after careful consideration that incorporates four data sets, Kaggle, McIntire, Reuters, and BuzzFeed, for two reasons. First of all, they both have a framework that is divided into two categories—real news and false news. Additionally, integrating the data sets lessens the shortcomings and bias of each separate data collection. Table III displays the WELFake open data set, which consists of 72 134 news items divided into 35 028 legitimate and 37 106 false reports . The data collection has three columns (title, text, and label), each of which has a binary label designating whether it is false or authentic news. The distribution of false and authentic news across all four feature categories in the WELFake data set is summarised in Table IV.

a) There are more short sentences (under ten words) that indicate true news than false news.

b) Fake news' text readability is worse than that of legitimate news.

d) Fake news stories are more subjective than legitimate news items.

d) The proportion of articles reporting actual news is higher than the proportion reporting false news.

2) Data Preprocessing: Depending on the data set and goals, this process addresses a variety of issues in the collected data, such as typographical errors, unstructured data format, and other limitations. a) Missing data addresses undefined (NaN) and blank values (NULL), which are present in the data set and impede the feature engineering process. We used a missing value imputation procedure, which guesses missing values and then analyses the whole data set as if these values were the real observed ones. This was done since removing the data entries containing missing values might result in the loss of crucial information.

b) Data that is inconsistent diverges from other data points as a result of errors made during data gathering. For outlier detection and correction, we used a variety of mathematical operations and visualisation approaches, including box plot, scatter plot, Z-score, and inter-quartile range (IQR) score.

c) Redundancy that might result in biassed findings, which may happen when the same individual gathered the data, is removed via duplicate data or deduplication.

d) Irrelevant data eliminates stop words (and other noise) that, although grammatically necessary, do not have any semantic value when it comes to news categorization processes.

Removing stop words while maintaining pertinent tokens improves the model's performance just somewhat.

e) To boost accuracy, stemming breaks down text into its constituent words using the Porter-Stemmer algorithm. If it is unable to identify the root word, it creates the canonical form of a related term.

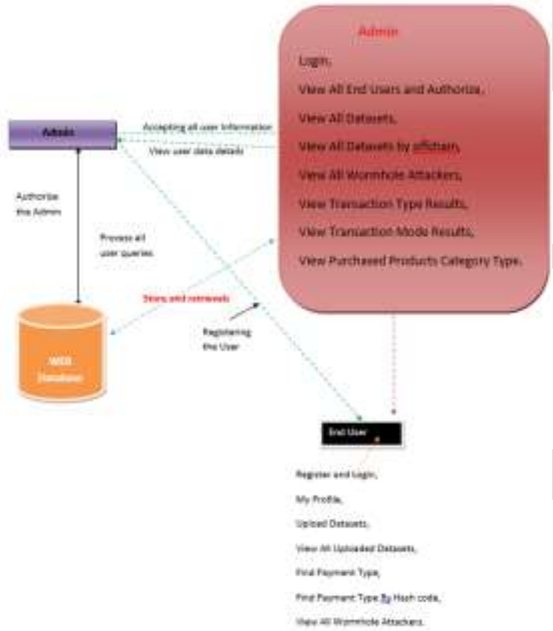

Fig 2 : System Architecture

3.4 Data Flow Diagram : Whenever a new system is developed, user training is required to educate them about the working of the system so that it can be put to efficient use by those for whom the system has been primarily designed. For this purpose the normal working of the project was demonstrated to the prospective users. Its working is easily understandable and since the expected users are people who have good knowledge of computers, the use of this system is very easy.

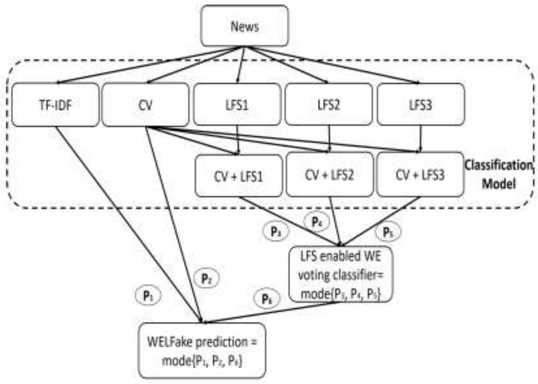

Fig 3: Data Flow Diagram

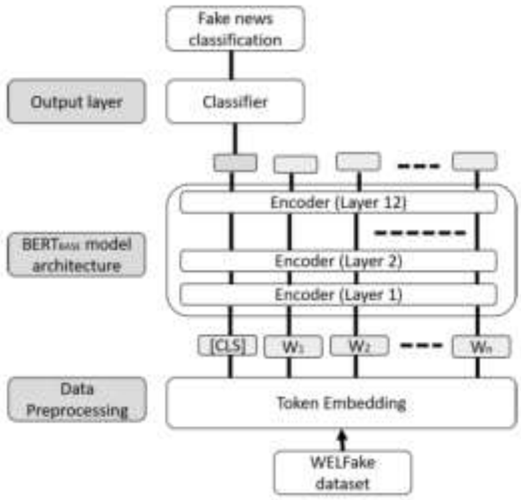

Fig 4 Use Case UML Diagrams

5. CONCLUSION

In order to identify bogus news in text, we developed a brand-new model called WELFake. In order to lessen the limitations and bias of the four open-source data sets (Kaggle, McIntire, Reuters, and BuzzFeed), we produced a bigger data collection called WELFake with approximately 72 000 news stories. After that, we examined over 80 linguistic variables from cutting-edge studies and chose 20 that were particularly important in order to reduce computing complexity and boost the precision of the conventional classifiers. We used six ML models (KNN, SVM, NB, DT, Bagging, and AdaBoost) to apply two WE-based approaches (TF-IDF, CV) across these linguistic characteristics and discovered that CV generates greater overall accuracy than TF-IDF using an SVM model. In order to classify the 20 variables, we employed CV rather than LFS and divided them into four groups: writing style, readability index, psycho-linguistics, and quantity.

We created three LFS by evenly distributing the twenty chosen characteristics among these categories since the voting classifier requires an odd number of predictions. Then, we applied all six ML models and incorporated the CV with these LFS. From each WE-enabled LFS data set, we selected the best accurate ML model and used its anticipated outcomes for voting categorization.

Finally, using the best model results from TF-IDF and CV over LFS, we applied the outcome of this voting classifier to the next level voting classification to produce the final classification.

The WELFake model provides a high 96.73% accuracy on the WELFake data set, according to experimental findings. We tested it with two cutting-edge works to further examine its advantages and discovered that it increases overall accuracy by 1.31% when compared to BERT and 4.25% when compared to CNN models. The accuracy on the McIntire and BuzzFeed data sets was also increased by up to 10% using the suggested WELFake model. SVM gave the most accurate results, according to our analysis of the performance of several ML models in terms of accuracy, precision, recall, and F1-score.

Finally, by up to 1.73%, our frequency-based model for examining writing patterns surpassed predictive-based comparable works applied using the Word2vec WE technique.

6. FUTURE SCOPE

In order to better verify the results produced by the WELFake model, we want to expand our research in the future to include other elements such as knowledge graphs and user credibility.

7. REFERENCES

[1] W. Jiang, J. Wu, F. Li, G. Wang, and H. Zheng, “Trust evaluation in online social networks using generalized network flow,” IEEE Trans. Comput., vol. 65, no. 3, pp. 952–963, Mar. 2016.
[2] M. Alrubaian, M. Al-Qurishi, A. Alamri, M. Al-Rakhami, M. M. Hassan, and G. Fortino, “Credibility in online social networks: A survey,” IEEE Access, vol. 7, pp. 2828–2855, 2019.
[3] S. Ranganath, S. Wang, X. Hu, J. Tang, and H. Liu, “Facilitating time critical information seeking in social media,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 10, pp. 2197–2209, Oct. 2017.
[4] Z. Zhang, R. Sun, X. Wang, and C. Zhao, “A situational analytic method for user behavior pattern in multimedia social networks,” IEEE Trans. Big Data, vol. 5, no. 4, pp. 520–528, Dec. 2019.
[5] M. Schudson and B. Zelizer, “Fake news in context,” in Understanding and Addressing the Disinformation Ecosystem. Philadelphia, PA, USA: Annenberg School for Communication, Apr. 2017, pp. 1–4.
[6] S. Zaryan, “Truth and trust: How audiences are making sense of fake news,” M.S. thesis, Media Commun. Studies, Lund Univ. Publications Student Papers, Stockholm, Sweden, Jun. 2017.
[7] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” Science, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.

[8] R. Sequeira, A. Gayen, N. Ganguly, S. K. Dandapat, and J. Chandra, “A large-scale study of the Twitter follower network to characterize the spread of prescription drug abuse tweets,” IEEE Trans. Comput. Social Syst., vol. 6, no. 6, pp. 1232–1244, Dec. 2019.

[9] Politifact News Dataset. Accessed: Mar. 31, 2020.

[10] The Washingtonpost Fact Checker. Accessed: Mar. 31, 2020.

[11] G.Sumalatha , K.Harish, M.Raghavaraju, R.Manideep, Rukhsar Jabeen “A Supervised Learning Estimator for Classifying Fake News Articles Using NLP To Identify in-Article Attribution”, , Journal of Interdisciplinary Cycle Research, Volume XIV, Issue III, March/2022 ISSN NO: 0022-1945